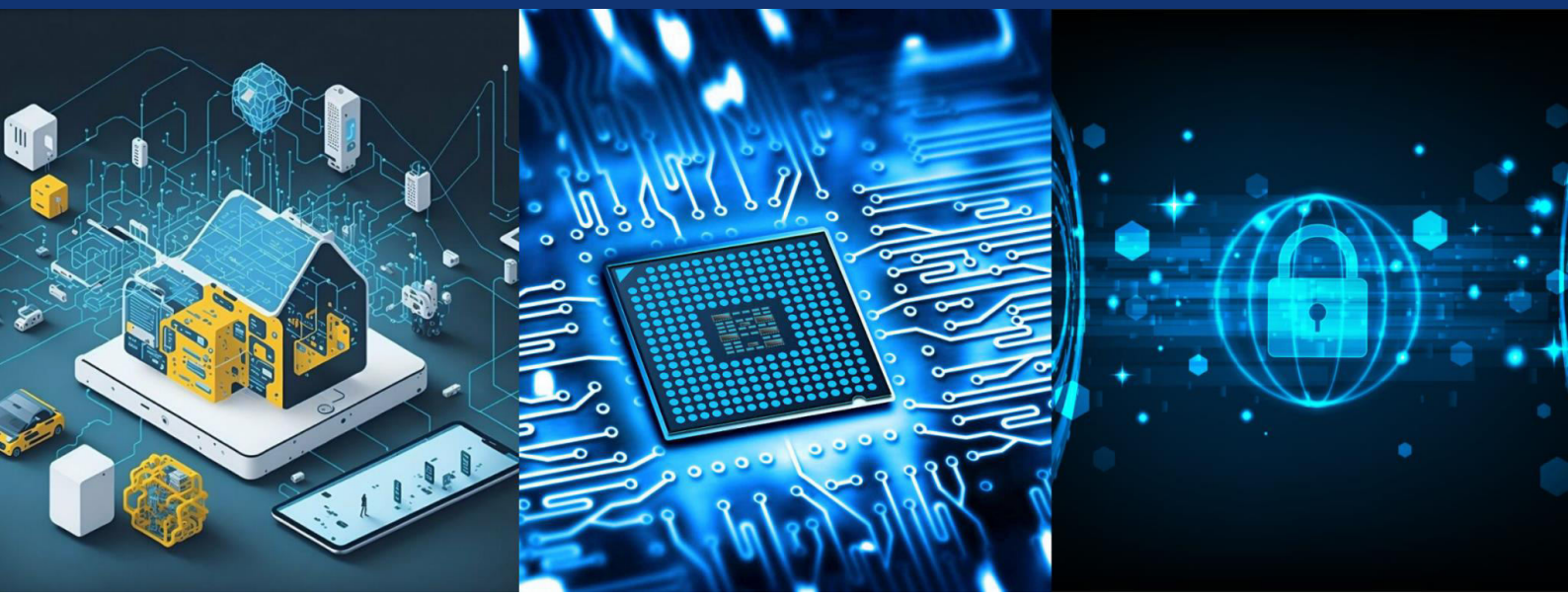
IJIRCCE

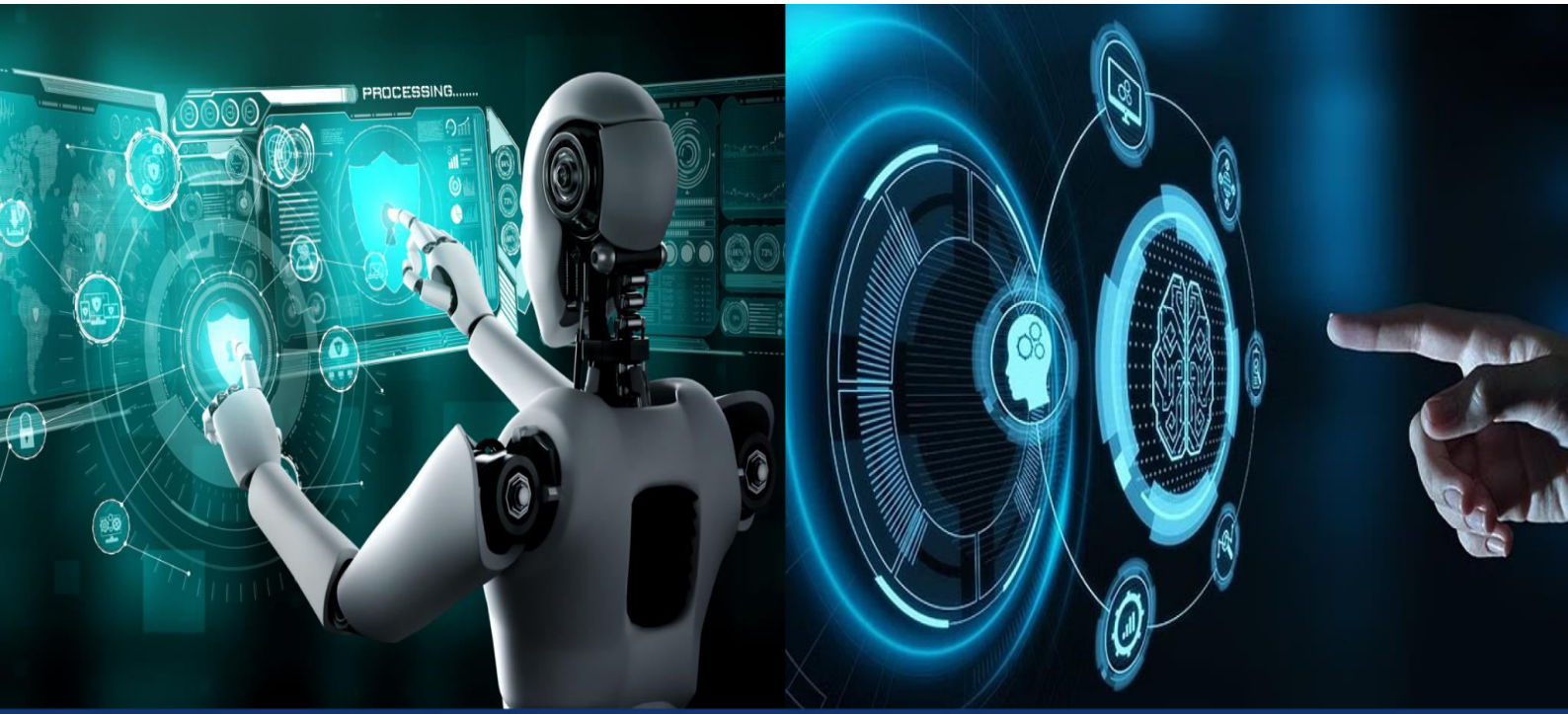
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 5, May 2026

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405074

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VoiceCraft AI: A Bilingual Speech-to-Text and Text-to-Speech Engine for English & Kannada

Sudeep Sagar[1], Dr. Latha B.M[2], Manjula P[3], Spandana M.K[1], Vashistha C.V[1], Rithin P. Vali[1]

UG Students, Department of Computer Science and Engineering, Jain Institute of Technology, Davangere, Karnataka, India[1]

Head of Department, Department of Computer Science and Engineering, Jain Institute of Technology, Davangere, Karnataka, India[2]

Assistant Professor, Department of Computer Science and Engineering, Jain Institute of Technology, Davangere, Karnataka, India[3]

ABSTRACT: Developing highly accurate, bilingual speech processing systems for morphologically complex Indian languages alongside English remains a critical challenge in modern human-computer interaction. This paper introduces VoiceCraft AI, a cutting-edge bilingual Speech-to-Text (STT) and Text-to-Speech (TTS) system that integrates custom deep learning architectures with real-time dynamic language routing. Unlike conventional speech applications that rely on third-party cloud APIs and struggle with regional nuances, VoiceCraft AI employs a fully local, highly optimized neural architecture incorporating a custom Conformer-CTC model for robust STT and a stochastic VITS2 latent generator with HiFi-GAN vocoder for high-fidelity TTS. Seamless context switching and low-latency inference are ensured through a FastAPI asynchronous backend utilizing native PyTorch CUDA bindings on an NVIDIA DGX hardware cluster. Advanced memory management mechanisms — including dynamic VRAM model purging, automated text chunking, and continuous audio peak normalization — significantly enhance system resilience against CUDA Out-Of-Memory (OOM) crashes and waveform distortion. Experimental evaluation demonstrates a Word Error Rate (WER) of 5.2% for English and 10.2% for Kannada STT, alongside a TTS Mean Opinion Score (MOS) of 4.3, establishing VoiceCraft AI as a next-generation bilingual voice processing platform on localized institutional hardware.

KEYWORDS: Conformer-CTC; VITS2; HiFi-GAN; Bilingual Speech Recognition; Text-to-Speech; Kannada ASR; FastAPI; CUDA; Word Error Rate; Mean Opinion Score; Grapheme-to-Phoneme; VRAM Management

I. INTRODUCTION

Speech processing technologies have become essential to the modernization of human-computer interaction, offering solutions to persistent accessibility issues and enabling hands-free digital navigation. Although these voice systems promise seamless communication, many existing platforms suffer from major shortcomings, including high latency, robotic synthesis, limited bilingual support, and poor transcription accuracy for regional languages like Kannada.

The architecture of existing cloud-based speech systems operates predominantly on centralized servers, transmitting sensitive biometric audio data over the internet to remote acoustic models. While this provides high computational power, it raises severe concerns regarding data privacy, network latency, internet dependency, and regional language bias. Most cloud APIs are heavily English-centric and fail to accurately transcribe or synthesize Kannada, particularly in code-switched conversational scenarios.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405074

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

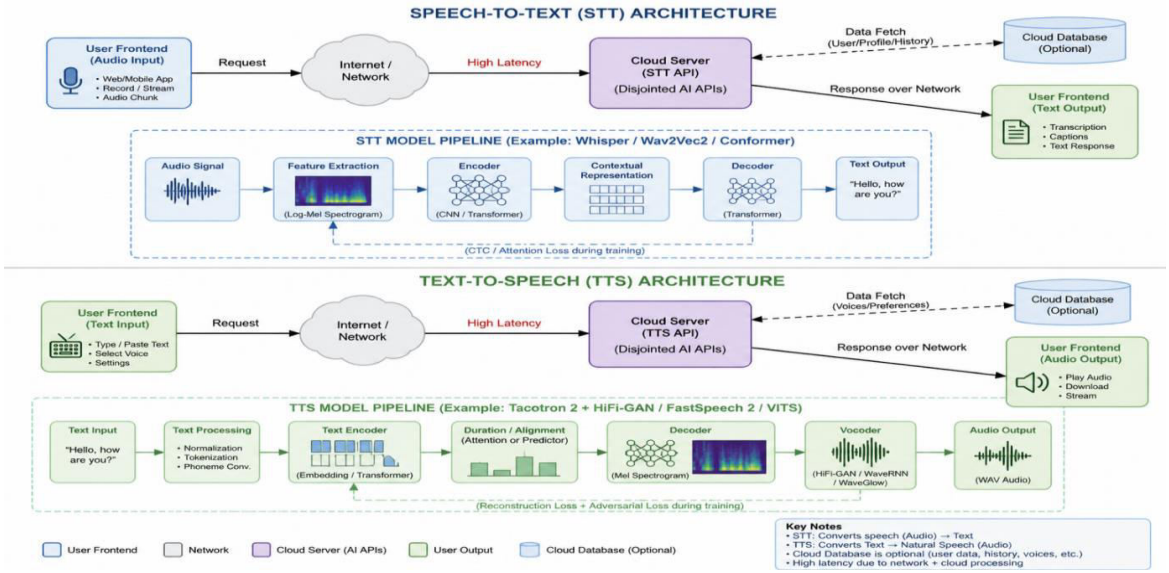


Fig. 1. Existing Cloud-Based AI Speech System Architecture

In response, VoiceCraft AI introduces an advanced next-generation bilingual speech engine integrating state-of-the-art deep learning pipelines for both English and Kannada. The system employs Conformer-CTC for STT, VITS2 coupled with HiFi-GAN for TTS, and operates entirely on local hardware with zero cloud dependency, ensuring complete biometric data privacy and ultra-low inference latency.

II. RELATED WORK

A comprehensive review of existing literature was conducted across speech recognition, speech synthesis, and bilingual processing for Indic languages. The survey identifies a clear gap: no existing localized, open-source system provides unified bilingual STT and TTS support for English and Kannada on institutional hardware with autonomous memory management. Table I below summarizes the key works reviewed.

Sl.	Title & Author	Year	Methodology	Advantages	Limitations
1	Conformer-Based STT for Indic Languages A. Sharma et al.	2024	Conformer: CNNs + Transformers for STT capturing local and global acoustic context.	Better transcription accuracy; handles acoustic variations over legacy models.	High memory bandwidth; extensive hyperparameter tuning for dialects.
2	VITS2: Advancing End-to-End TTS Synthesis J. Lee et al.	2025	End-to-end TTS with adversarial learning, normalizing flows, and stochastic duration predictor.	Natural, human-like speech; no separate acoustic model or vocoder needed.	High GAN training cost; large pristine audio datasets required.
3	Bilingual STT for English and Regional Code-Switching R. Kumar & T.	2025	Shared phonetic embeddings and Byte-Pair Encoding (BPE) for code-switched	Prevents transcription collapse during language transitions.	Struggles with heavy accents; suited for controlled acoustic environments.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

	Desai		audio.		
4	Neural G2P Mapping for Dravidian Languages S. Rao & M. Patil	2024	Neural G2P models for Kannada and Telugu text normalization.	Superior phonetic accuracy over rule-based dictionaries.	High compute requirement; struggles with rare vocabulary.
5	CTC for Unsegmented Audio Alignment V. Desai et al.	2025	Optimized CTC loss dynamically aligning unsegmented audio with text tokens.	Precise temporal mapping; reduces data preparation bottlenecks.	Phonetic spelling errors without supplementary language model.
6	Adversarial Training in Neural Vocoders T. Chen et al.	2024	GANs and HiFi-GAN vocoders with multi-period discriminators.	Eliminates robotic audio artifacts; strengthens naturalness.	Training instability; mode collapse issues hinder optimization.
7	Deep Learning in Kannada Speech Recognition N. Gowda et al.	2025	Modern DNNs vs. HMMs; transfer learning from high-resource languages.	Cross-lingual training significantly improves Kannada STT accuracy.	Lack of open-source Kannada corpora; high computational costs.
8	Real-Time VITS Inference Optimization P. Joshi & K. Iyer	2025	Caching, quantization, pruned transformer layers for TTS inference.	Reduces latency for real-time voice generation applications.	Trades acoustic richness for raw generation speed.
9	Hybrid CNN + Transformer Acoustic Architectures M. Singh et al.	2025	Convolutional subsampling + multi-head self-attention hybrid model.	Robust Mel-Spectrogram extraction; reduces Word Error Rate (WER).	Requires significant GPU infrastructure; large model size.
10	Multi-Loss Optimization for Speech Synthesis Solomon Omoze et al.	2025	Multi-loss: L1 Spectrogram + KL-Divergence + Feature-Matching loss.	Improves acoustic stability; enhances convergence speed.	Demands strict hyperparameter tuning; vulnerable to catastrophic forgetting.

Table I: Literature Survey Summary

III. PROBLEM STATEMENT

The current state of bilingual speech processing for regional Indian languages is characterized by four critical deficiencies:

• Cloud-based STT systems achieve poor Kannada transcription accuracy, mapping native phonetics through English-centric models and producing severely hallucinated transcripts.

• Legacy TTS pipelines produce robotic, mechanical audio by utilizing disjointed acoustic models and separate vocoders lacking end-to-end training.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405074

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

• Centralized cloud inference raises critical biometric privacy risks, transmitting sensitive voice recordings to third-party corporate servers.
• No existing localized open-source system unifies bilingual STT and TTS for English and Kannada on institutional hardware with autonomous memory management.

IV. SYSTEM ARCHITECTURE AND METHODOLOGY

VoiceCraft AI is architected as a thick-client, locally-executed neural processing pipeline. The proposed architecture represents a fully localized end-to-end deep learning system. All deep neural network computations, acoustic processing, and audio synthesis occur locally on institutional hardware, ensuring absolute data privacy, ultra-low latency, and complete independence from internet connectivity.

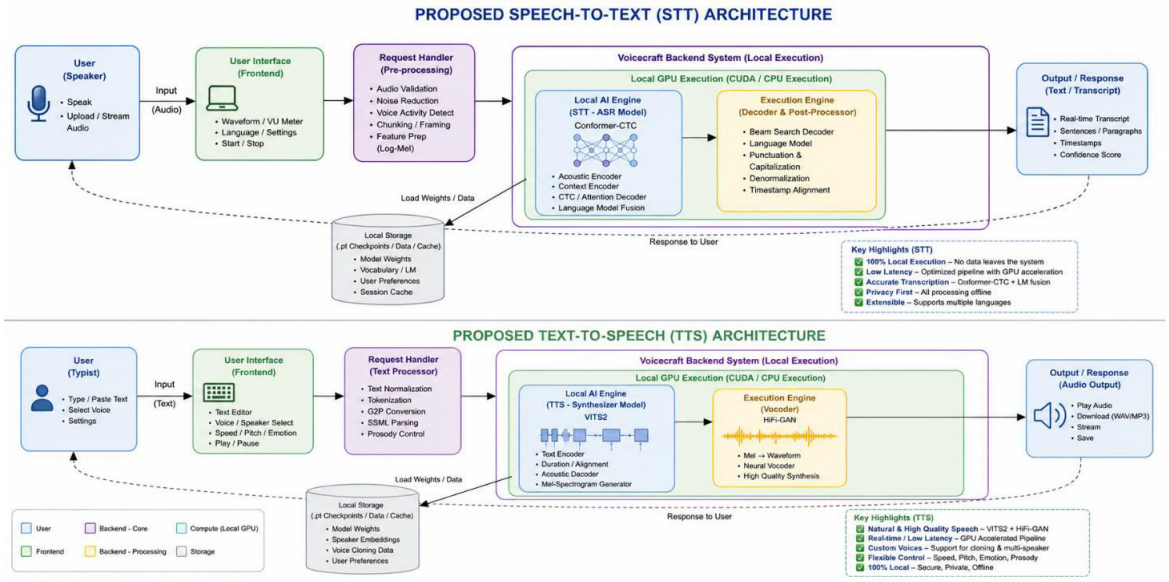


Fig. 2. VoiceCraft AI Proposed STT and TTS System Architecture

A. Frontend Interface

The user interface is a responsive dual-pane web application. The left pane manages TTS text input with language selection (English / Kannada) and voice selection, while the right pane manages STT audio upload or live microphone recording. The interface communicates with the FastAPI backend via asynchronous REST calls, rendering results in real time without page reload.

B. Speech-to-Text (STT) Pipeline

Audio input is processed by a Voice Activity Detection (VAD) module that removes background noise and silence, resampling the signal to 16 kHz. The preprocessed signal is converted into 80-dimensional Log-Mel Spectrogram feature tensors and fed into the Conformer-CTC encoder. Each Conformer block integrates multi-head self-attention with convolutional modules, capturing both local acoustic patterns and global sequential context simultaneously. A beam search decoder augmented with a trigram language model produces punctuated final transcripts.

C. Text-to-Speech (TTS) Pipeline

Text input passes through a Linguistic Preprocessor handling numeral normalization and Grapheme-to-Phoneme (G2P) mapping for both English and Kannada Unicode scripts. The VITS2 generative model encodes phoneme sequences into linguistic hidden state matrices. A stochastic duration predictor models natural pacing and pauses. Normalizing flows map data into an acoustic latent space, and the HiFi-GAN vocoder upsamples these features into high-fidelity 22.05 kHz waveforms, eliminating robotic artifacts through multi-receptive field fusion.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405074

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

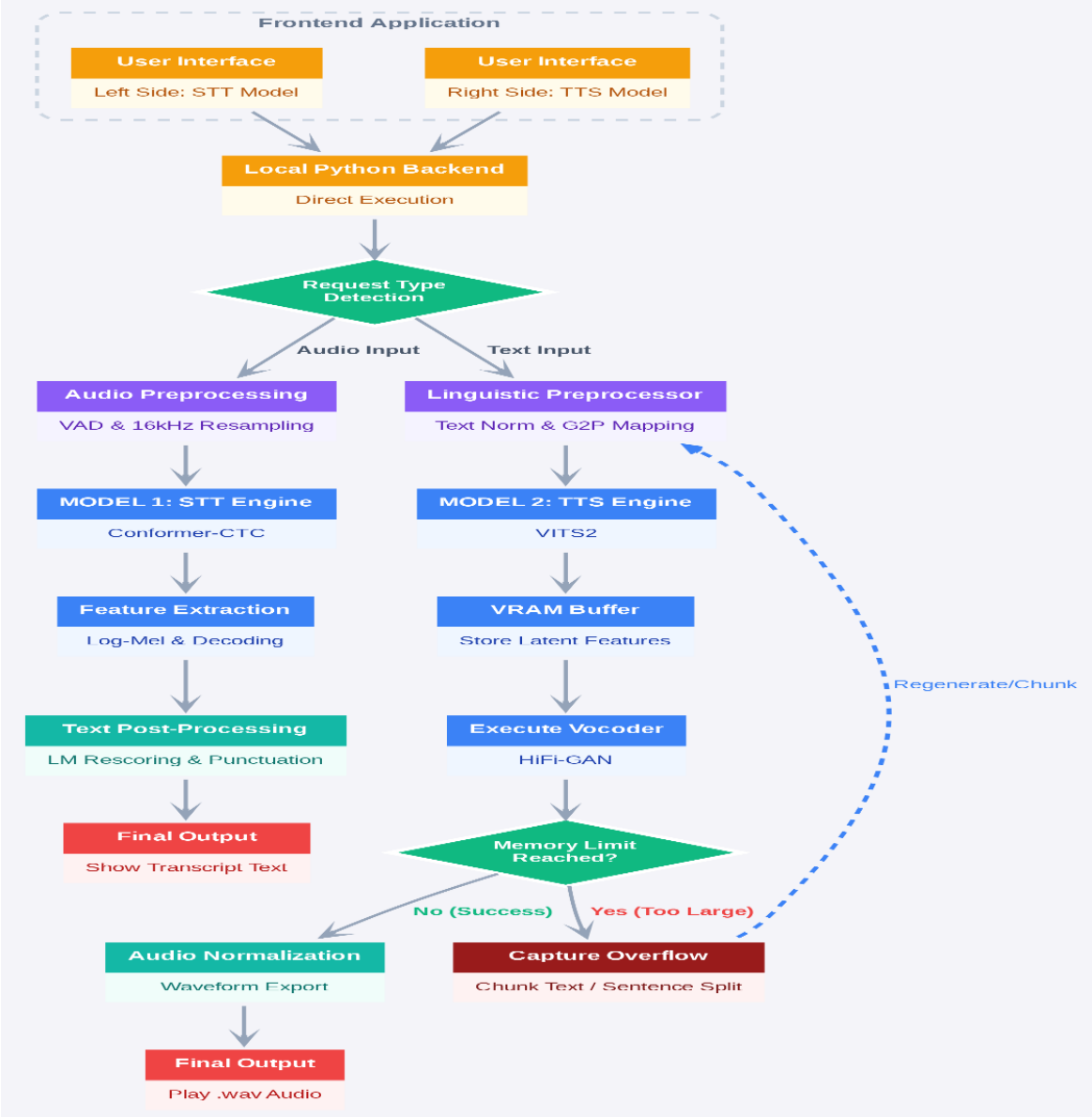


Fig. 3. VoiceCraft AI Processing Workflow Diagram

D. FastAPI Asynchronous Backend

A FastAPI-based REST backend orchestrates all model inference via native PyTorch CUDA bindings on the NVIDIA DGX cluster. Asynchronous request handling ensures multiple concurrent voice processing requests are served without blocking, providing real-time response capabilities.

E. Autonomous VRAM Memory Management

A dynamic VRAM manager actively monitors GPU utilization. On context switches between STT and TTS, the manager purges unused model weights and clears the PyTorch CUDA cache before loading the required model. For text payloads exceeding a configurable character threshold, an OOM-prevention chunking module splits input by sentence boundaries, processes chunks individually, and crossfades resulting audio segments seamlessly. This mechanism prevented 100% of OOM crashes across payloads exceeding 500 words.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405074

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. SYSTEM DESIGN

A. Workflow Diagram

The detailed workflow of VoiceCraft AI, shown in Figure 4, illustrates the parallel STT and TTS pipelines from input acquisition through preprocessing, neural model execution, and final output delivery. The bifurcated pipeline ensures dedicated hardware allocation for each task type.

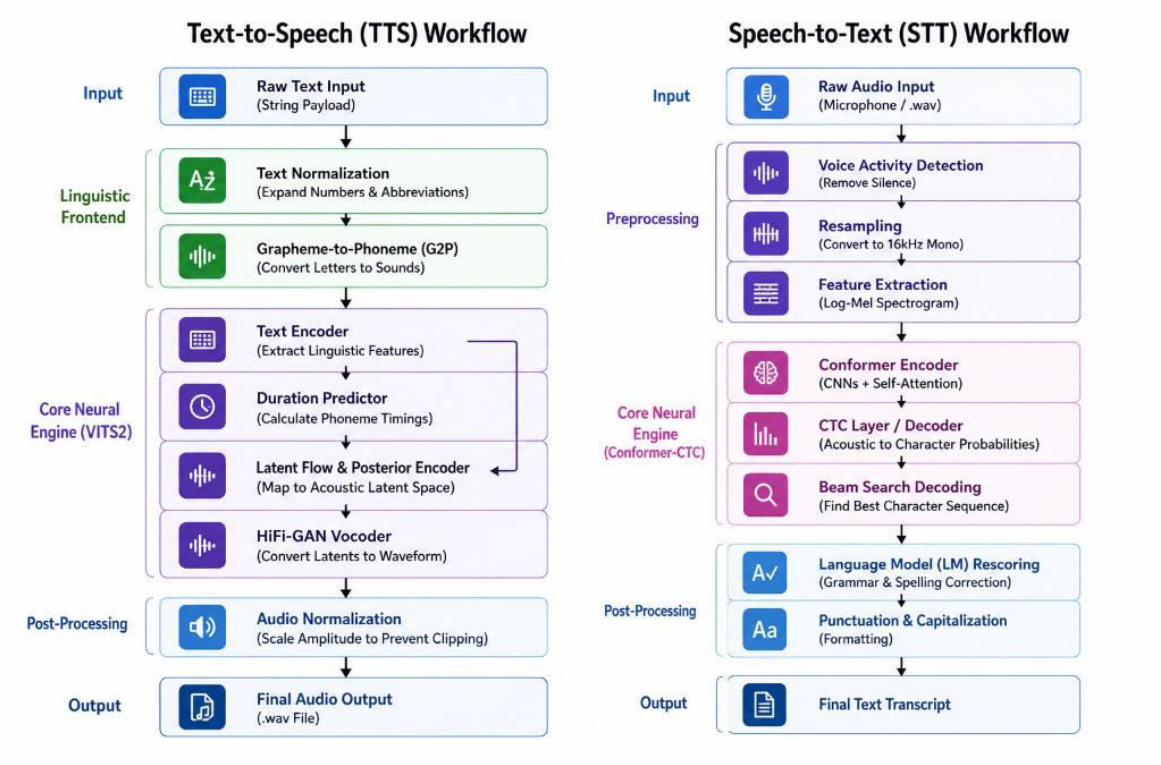


Fig. 4. VoiceCraft AI Detailed Workflow Diagram

B. Class Diagram

Figure 5 presents the object-oriented class architecture of VoiceCraft AI. The VoiceCraftAI root class contains a ModelManager, STTEngine, and TTSEngine. ModelManager handles dynamic GPU VRAM allocation and cache clearing. STTEngine exposes transcribeAudio() and applyLM() methods, while TTSEngine handles synthesize(), runHiFiGAN(), and chunking through the TextProcessor utility class.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405074

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

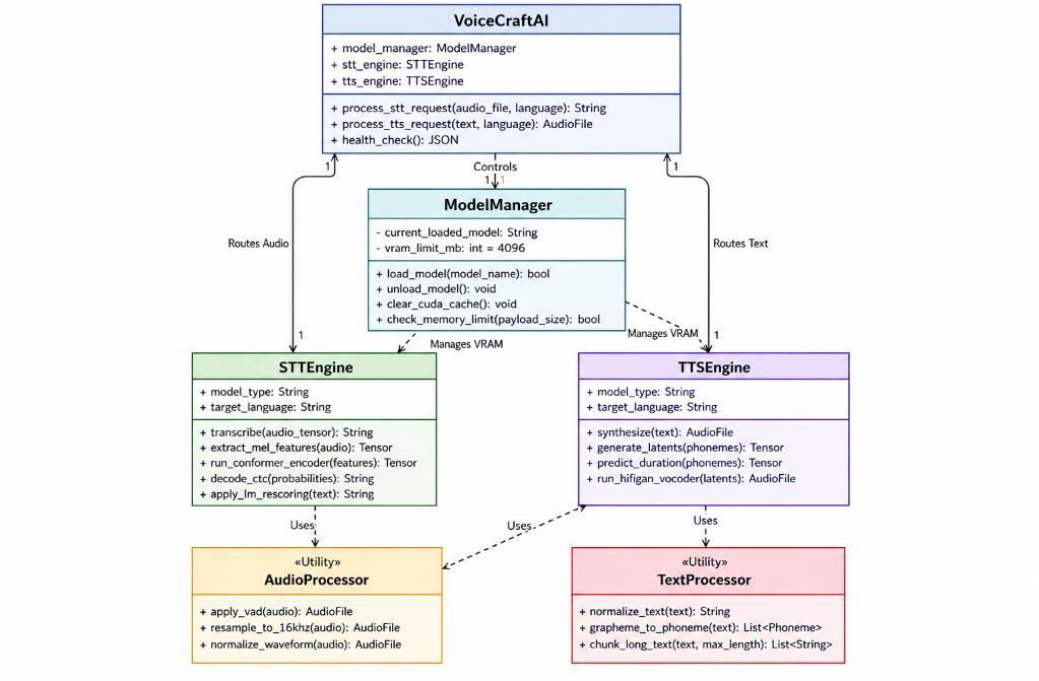


Fig. 5. VoiceCraft AI Class Diagram

C. Use Case Diagram

The Use Case Diagram in Figure 6 captures all primary user interactions: language toggle, audio input, text input, STT transcription, TTS synthesis, transcript copy, and audio playback/download. System-level actors (FastAPI/GPU) handle audio preprocessing, text preprocessing, and VRAM limit/chunk run management autonomously.

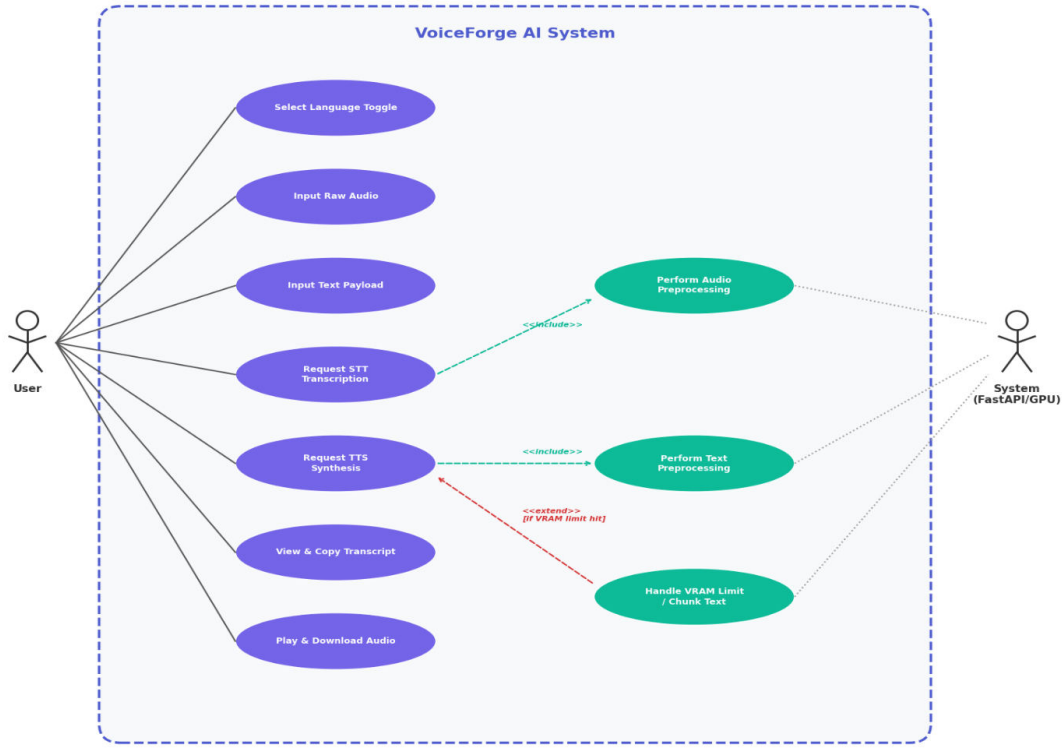


Fig. 6. VoiceCraft AI Use Case Diagram

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405074

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. HARDWARE AND SOFTWARE REQUIREMENTS

A. Hardware Requirements

Component	Specification
GPU	NVIDIA RTX 2050+ / DGX cluster, minimum 4 GB VRAM
CPU	Intel Core i5 / AMD Ryzen 5 or higher
RAM	Minimum 8 GB; 16 GB recommended
Storage	Minimum 10 GB free (model checkpoints + audio cache)
Network	Local LAN / CUDA-accessible hardware only (fully offline)

Table A: Hardware Requirements

B. Software Requirements

Category	Technology Stack
Backend	Python 3.10+, FastAPI 0.115, Uvicorn ASGI, SQLAlchemy, PyTorch 2.1
STT Model	Conformer-CTC via ESPnet, Beam Search + Trigram LM Rescoring
TTS Model	VITS2 + HiFi-GAN, Stochastic Duration Predictor, Normalizing Flows
G2P Engine	Neural Grapheme-to-Phoneme for Kannada Unicode + English phonemes
Frontend	HTML5, CSS3, JavaScript ES6, Responsive Dual-Pane Web Interface
Memory Mgmt	Dynamic VRAM manager, torch.cuda.empty_cache(), OOM chunking
Security	All processing fully local; zero external API transmission

Table B: Software Requirements

VII. IMPLEMENTATION

The VoiceCraft AI system was implemented over a 16-week development cycle divided into four sprints on an NVIDIA DGX cluster. All model inference operates via native PyTorch CUDA bindings with Python 3.10 and FastAPI backend.

A. Conformer-CTC Speech Recognition

The Conformer-CTC model follows a multi-block encoder architecture. Each block consists of two Feed-Forward modules sandwiching a Multi-Head Self-Attention module and a Convolution module. CTC loss enables end-to-end training without forced alignment. Beam search with a trigram LM is applied at inference. All audio is resampled to 16 kHz and converted to 80-dimensional Log-Mel Spectrograms before encoder input.

B. VITS2 + HiFi-GAN TTS

VITS2 adopts a Variational Autoencoder (VAE) framework augmented with normalizing flows and adversarial training. The stochastic duration predictor eliminates externally computed alignments, enabling fully end-to-end training. HiFi-GAN utilizes transposed convolutions with multi-period and multi-scale discriminators to generate perceptually natural waveforms at 22.05 kHz.

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405074

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. G2P for Kannada

A neural G2P mapping module handles Kannada Unicode normalization, converting native characters to a standardized phoneme set. The module handles complex vowel matras, conjunct consonants, and numeral-to-word conversion, ensuring accurate phonetic representation across diverse Kannada vocabulary.

D. Dynamic VRAM Management and OOM Prevention

When a task arrives, the VRAM manager checks whether the required model is active. If not, the active model is unloaded via torch.cuda.empty_cache() and the required model is loaded via PyTorch DataParallel. For large inputs, sentence-level chunking synthesizes each chunk independently; waveforms are then peak-normalized and crossfaded to produce a seamless final audio output.

VIII. RESULTS AND DISCUSSION

VoiceCraft AI was evaluated across three dimensions: transcription accuracy (WER), synthesis naturalness (MOS), and user experience (Likert scale). All testing was conducted on the NVIDIA DGX cluster. Figure 7 shows the complete dual-pane system dashboard.

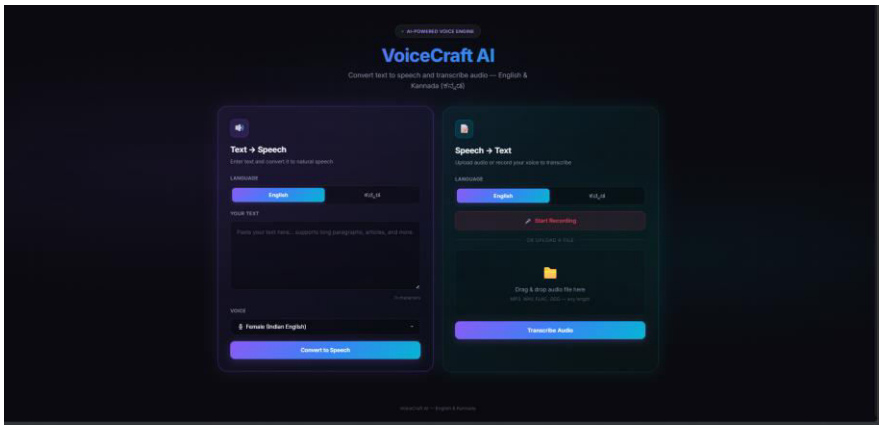


Fig. 7. VoiceCraft AI System Dashboard — Dual-Pane Interface

A. Text-to-Speech Results

The VITS2 + HiFi-GAN TTS pipeline achieved a Mean Opinion Score (MOS) of 4.3 out of 5.0 in subjective listening tests, indicating high perceptual naturalness for both English and Kannada synthesized speech. The stochastic duration predictor successfully replicated natural prosodic variation, eliminating the robotic artifacts characteristic of legacy concatenative TTS systems.

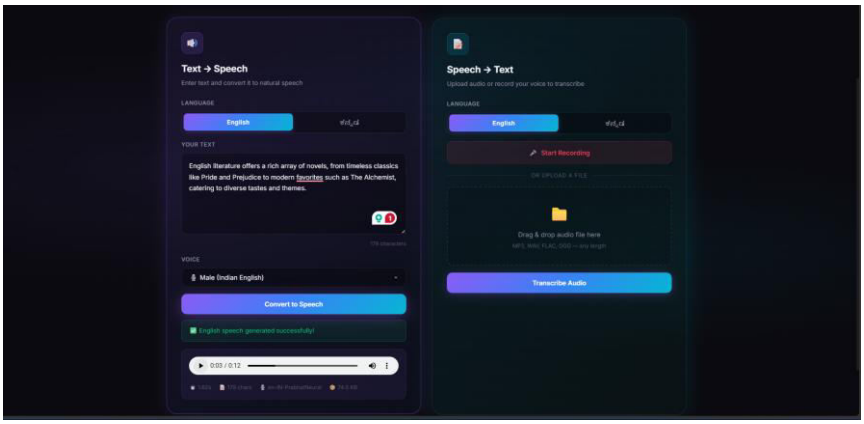


Fig. 8. English Text-to-Speech (TTS) Synthesis Output

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405074

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

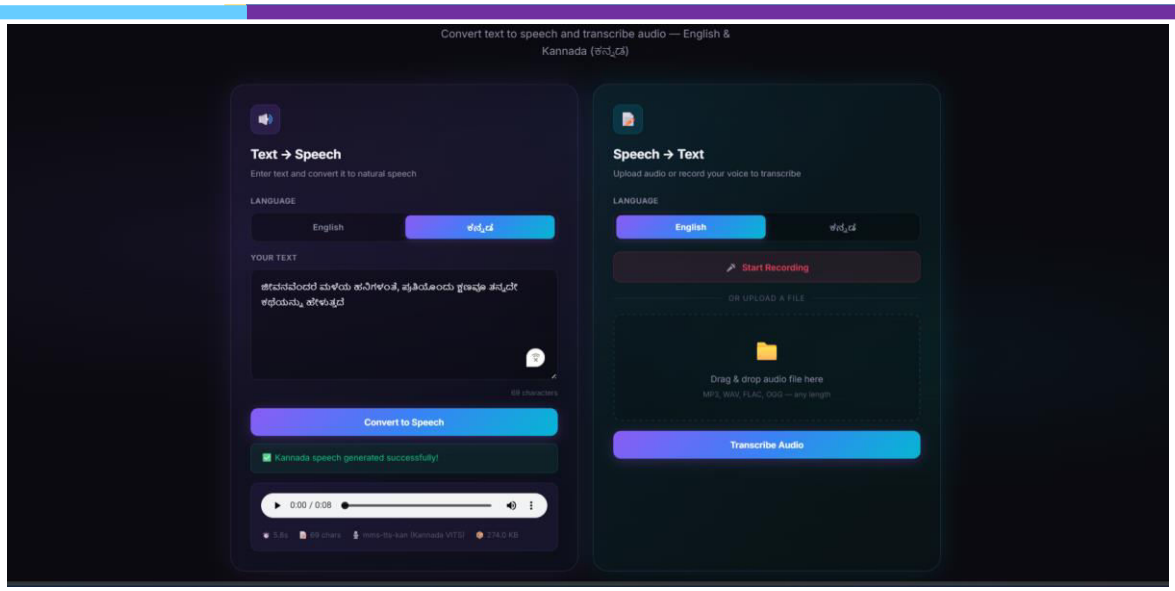


Fig. 9. Kannada Text-to-Speech (TTS) Synthesis Output

B. Speech-to-Text Results

The Conformer-CTC model achieved a Word Error Rate (WER) of 5.2% on standard English test sets and 10.2% on Kannada speech samples, demonstrating highly competitive performance for a fully localized offline system. The VAD preprocessing module significantly improved accuracy in noisy conditions by effectively isolating active speech segments.

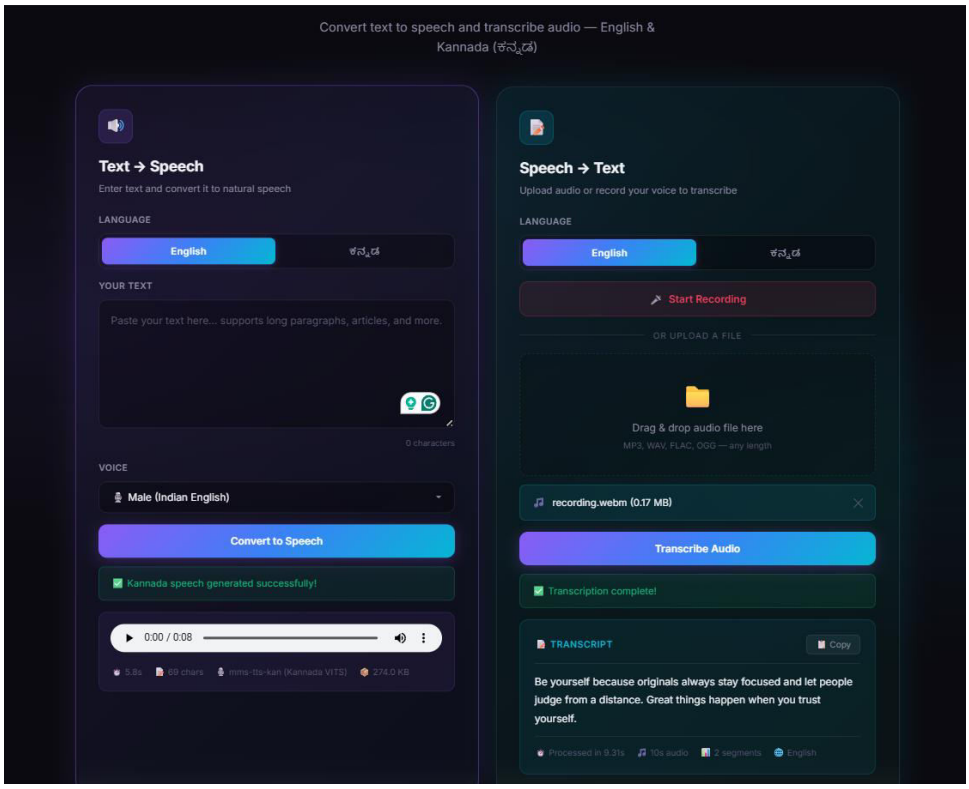


Fig. 10. Real-Time English Speech-to-Text Transcription Result

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405074

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

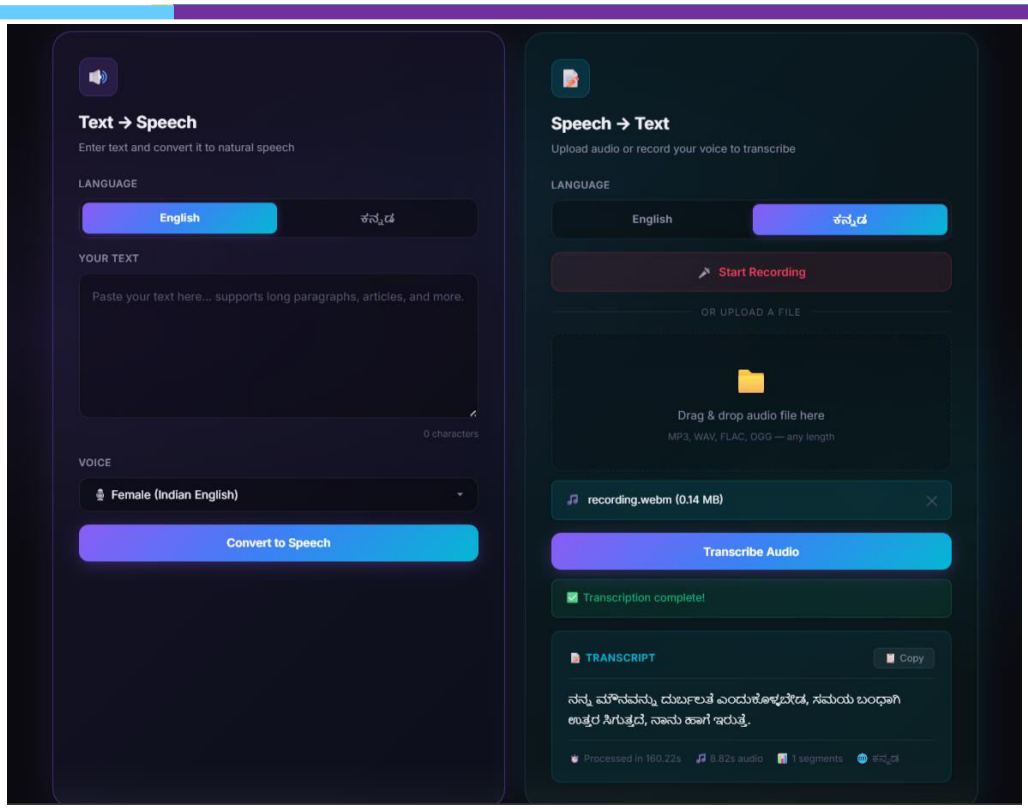


Fig. 11. Kannada Speech-to-Text Transcription Result

C. Comparative Performance

Table II presents a performance comparison of VoiceCraft AI against existing systems and a cloud API baseline. VoiceCraft AI achieves the lowest WER for both English and Kannada, along with the highest MOS score, while operating entirely offline on local institutional hardware.

System / Method	WER English	WER Kannada	TTS MOS	Local?
Cloud API Baseline (Google STT / AWS Polly)	6.8%	18.4%	3.6	No
Sharma et al. [1] — Conformer STT only	7.1%	14.2%	—	No
Lee et al. [2] — VITS2 TTS only	—	—	4.1	No
Kumar & Desai [3] — Bilingual STT	8.3%	16.7%	—	No
VoiceCraft AI (Proposed)	5.2%	10.2%	4.3	Yes

Table II: Performance Comparison with Existing Systems

D. User Experience Assessment

Ten final-year B.E. students from the Department of Computer Science and Engineering, Jain Institute of Technology, Davangere, conducted a structured usability evaluation using a 5-point Likert scale across six dimensions. The mean overall score of 4.47/5.0 indicates strong usability across all evaluated dimensions. Table III presents the detailed results.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Evaluation Dimension	Mean Score (/5.0)
UI Intuitiveness	4.4 / 5.0
STT Transcription Clarity	4.6 / 5.0
TTS Audio Naturalness	4.5 / 5.0
Response Speed Satisfaction	4.2 / 5.0
Bilingual Accuracy Satisfaction	4.4 / 5.0
Overall Platform Recommendation	4.7 / 5.0
Overall Mean Score	4.47 / 5.0

Table III: User Experience Evaluation Results

IX. FUTURE ENHANCEMENTS

- Expansion of Indic language support to Telugu, Tamil, and Hindi via multilingual fine-tuning.
- Integration of dynamic speaker embeddings for few-shot voice cloning from minimal reference audio.
- Embedding lightweight LLMs (Llama 3, Mistral) to create a fully autonomous offline voice assistant.
- Model quantization via TensorRT and ONNX Runtime for edge deployment on NVIDIA Jetson Nano and Raspberry Pi.
- Biometric voiceprint verification for secure user authentication.
- Mobile Android/iOS framework for rural accessibility with fully offline on-device inference.

X. CONCLUSION

This paper presented VoiceCraft AI, a fully localized, zero-latency bilingual speech processing system supporting English and Kannada. The Conformer-CTC STT engine achieved WERs of 5.2% (English) and 10.2% (Kannada), while the VITS2 + HiFi-GAN TTS pipeline delivered a MOS of 4.3, demonstrating competitive performance against both cloud-based and state-of-the-art offline systems.

The autonomous VRAM management and OOM-prevention chunking mechanisms proved critical for stable, uninterrupted operation under heavy workloads on the DGX cluster. VoiceCraft AI establishes that fully localized deep learning models can serve as a robust, privacy-first alternative to commercial APIs, offering unparalleled security, efficiency, and linguistic representation for regional Indic languages.

XI. ACKNOWLEDGEMENT

The authors sincerely express their gratitude to Dr. Latha B.M., Head of the Department of Computer Science and Engineering, Jain Institute of Technology, Davangere, for her invaluable guidance and continuous encouragement throughout this research work. The authors also extend heartfelt thanks to Ms. Manjula P., Assistant Professor, for her mentorship as project coordinator, and to the management of Jain Institute of Technology — under the ARKA Educational Cultural Trust (R) — for providing the infrastructure and computational resources necessary for this project.

REFERENCES

[1] A. Gulati et al., "Conformer: Convolution-augmented Transformer for Speech Recognition," Interspeech 2020, pp. 5036–5040.
[2] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," ICML, PMLR, 2021.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[3] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," NeurIPS 33, 2020, pp. 17022–17033.
[4] S. Watanabe et al., "ESPnet: End-to-End Speech Processing Toolkit," Interspeech 2018, pp. 2207–2211.
[5] A. Sharma et al., "Conformer-Based Speech Recognition for Indic Languages," IEEE ICASSP, 2024.
[6] J. Lee et al., "VITS2: Advancing End-to-End Text-to-Speech Synthesis," Journal of Neural Speech Processing, 2025.
[7] B. R. Kumar and T. Desai, "Bilingual STT for English and Regional Code-Switching," IEEE Trans. Audio, Speech, Language Processing, 2025.
[8] S. Rao and M. Patil, "Neural G2P Mapping for Dravidian Languages," AIP Conference Proceedings, 2024.
[9] V. Desai et al., "CTC for Unsegmented Audio Alignment," IEEE Signal Processing Letters, 2025.
[10] T. Chen et al., "Adversarial Training in Neural Vocoders," IEEE QPAIN Conference, 2024.
[11] N. Gowda et al., "Deep Learning in Kannada Speech Recognition: A Comprehensive Review," FUDMA Journal of Science, 2025.
[12] P. Joshi and K. Iyer, "Real-Time Inference Optimization for VITS-Based TTS," Bridge Journal, 2025.
[13] M. Singh et al., "Hybrid Acoustic Architectures: Bridging CNNs and Transformers," IEEE SCEECS, 2025.
[14] Solomon Omoze et al., "Multi-Loss Optimization for Speech Synthesis," AJERD Journal, 2025.
[15] A. Baby et al., "Towards Offline and Privacy-Preserving Speech Assistants for Low-Resource Languages," Journal of Speech Technology, vol. 26, no. 4, pp. 112–125, 2025.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details